REVIEW ARTICLE

MEDICAL EDUCATION

Educational Strategies for Clinical Supervision of Artificial Intelligence Use

Raja-Elie E. Abdulnour, M.D.,[1] Brian Gin, M.D., Ph.D.,[2] and Christy K. Boscardin, Ph.D.[3,4]

Author affiliations are listed at the end of the article. Dr. Abdulnour can be contacted at rabdulnour@bwh.harvard.edu or at the Department of Medicine, Brigham and Women's Hospital, 75 Francis St., Boston, MA 02115.

N Engl J Med 2025;393:786-97.
DOI: 10.1056/NEJMra2503232
Copyright © 2025 Massachusetts Medical Society.

Human–computer interactions have been occurring for decades, but recent technological developments in medical artificial intelligence (AI) have resulted in more effective and potentially more dangerous interactions. Although the hype around AI resonates with previous technological revolutions, such as the development of the Internet and the electronic health record,[1] the appearance of large language models (LLMs) seems different. LLMs can simulate knowledge generation and clinical reasoning with humanlike fluency, which gives them the appearance of agency and independent information processing.[2] Therefore, AI has the capacity to fundamentally alter medical learning and practice.[3,4] As in other professions,[5] the use of AI in medical training could result in professionals who are highly efficient yet less capable of independent problem solving and critical evaluation than their pre-AI counterparts.

Such a challenge presents educational opportunities and risks. AI can enhance simulation-based learning,[6] knowledge recall, and just-in-time feedback[7] and can be used for cognitive off-loading of rote tasks. With cognitive off-loading, learners rely on AI to reduce the load on their working memory, a strategy that facilitates mental engagement with more-demanding tasks.[8] However, off-loading of complex tasks, such as clinical reasoning and decision making, can potentially lead to automation bias (overreliance on automated systems and risk of error), "deskilling" (loss of previously acquired skills), "never-skilling" (failure to develop essential competencies), and "mis-skilling" (reinforcement of incorrect behavior due to AI errors or bias).[9] These risks are especially troubling because LLMs operate as unpredictable black boxes[10]; they generate probabilistic responses with low reasoning transparency, which limits assessment of their reliability. For example, in one study, more than a third of advanced medical students missed erroneous LLM answers to clinical scenarios.[11]

The inherent variability and potential inaccuracies of AI-generated output can leave even experienced clinicians uncertain about AI recommendations. This dilemma is not novel; it mirrors the broader challenge of confronting unfamiliar clinical problems. Such moments require adaptive practice — the capacity to shift fluidly between efficient, familiar, routinized behavior and innovative, flexible problem solving.[12] Critical thinking is the structured cognitive tool set that underlies adaptive practice in the face of uncertainty. It enables clinicians to bring assumptions to the surface and engage in self-reflection that helps them recognize knowledge gaps and biases, mitigate errors, adapt to new problems, and generate or adopt new knowledge (i.e., learn).[13,14] Thus, critical thinking is foundational to adaptive practice in the age of AI.

Clinicians supervising medical learners, henceforth referred to as educators, must explicitly teach, assess, and model critical thinking to promote lifelong adaptive

The New England Journal of Medicine is produced by NEJM Group, a division of the Massachusetts Medical Society.
Downloaded from nejm.org by a a on August 20, 2025.
Copyright © 2025 Massachusetts Medical Society. All rights reserved, including those for text and data mining, AI training, and similar technologies.

SHOUT
SHARE
IT

KEY POINTS

EDUCATIONAL STRATEGIES FOR CLINICAL SUPERVISION OF ARTIFICIAL INTELLIGENCE USE

- Use of artificial intelligence (AI) for the development of expert practice presents unprecedented opportunities but also poses risks, such as "deskilling," "never-skilling," and "mis-skilling."
- Clinical supervisors may be less experienced with AI than learners are. Faculty development should embrace shared learning environments that allow coexploration of AI capabilities and limitations.
- Adaptive practice — shifting between efficiency and innovation — is foundational in AI-enabled learning. Critical thinking supports this shift and must be taught and modeled.
- AI interactions lead to moments when clinicians receive outputs they cannot fully retrace, which prompts a leap of faith. Pausing to recognize these moments is essential for critical thinking.
- DEFT-AI (diagnosis, evidence, feedback, teaching, and recommendation for AI use) is a structured framework to promote critical thinking and AI literacy during learner–AI interactions.
- Two AI use behaviors emerge: cyborg (tight intertwining of user and AI for each task) and centaur (division of tasks between user and AI, with critical oversight). Adaptive AI practice requires the ability to shift between these behaviors according to the complexity of the task and the risk involved.

practice. Stakeholders are developing system-level strategies for safe AI integration in medical education,[15] but a critical gap remains: the absence of structured strategies that equip educators and learners with the necessary skills to engage critically with AI.[16] In this review, we propose a framework that leverages educational strategies to teach and assess critical thinking during clinical supervision of trainees wherever AI is being used.

TEACHING WHILE LEARNING

As AI tools permeate classrooms and clinical settings, educators find themselves supervising the use of a technology that learners may be more adept at using than the educators themselves are. This inversion of expertise parallels earlier shifts in medical education, such as the rise of the patient-centered medical home, where faculty had to teach, learn, and practice system change simultaneously.[17] In such contexts, faculty development must be grounded in a shared learning model by expanding the definition of educator to include all members of the clinical team (including AI-literate learners and patients) and supporting reflective, team-based "communities of practice."[17] These principles apply directly to AI supervision: educators should embrace moments of learner-led insight and invite shared inquiry into the capabilities and limitations of AI. Doing so transforms discomfort into an opportunity for comanagement of uncertainty,[18] which sets the stage for structured educational moments that promote clinical thinking and AI literacy for all. The strategies we describe below are not only tools for teaching but also scaffolds for educators to develop their own understanding of AI. In this new terrain, teachers are learners, too.

PROMISES AND PERILS OF AI IN MEDICAL LEARNING

Consider this fictitious but realistic example: during a clinic session, a medical resident discreetly consults his smartphone after evaluating a patient and prompts an LLM to generate a differential diagnosis and management plan (Fig. 1). Within seconds, the AI delivers a well-reasoned and persuasive argument. The learner inserts the AI-generated recommendations in the patient's note. Observing this interaction across the room, the educator thinks, "Now what? What prompts did the resident provide the AI? Is the resident questioning the AI or accepting its suggestions as they are? Can this AI be entrusted with clinical reasoning? Should I intervene, and if so, how? Is this the future of clinical reasoning?"

The educator's last question highlights the uncertainty around defining the use of AI to assist clinicians.[19] Today, AI interactions are occurring throughout a learner's workday. For example, learners with limited expertise in interpreting electrocardiographic (ECG) tracings often lean on the interpretation from the ECG machine to drive their clinical assessment of a patient with chest pain.[20] As technology evolves, clinicians may off-load many clinical tasks to AI, from image interpretation in radiology[21,22] to automated clinical documentation.[23]

The educator's question about whether to entrust AI with clinical reasoning is familiar and

The New England Journal of Medicine is produced by NEJM Group, a division of the Massachusetts Medical Society.
Downloaded from nejm.org by a a on August 20, 2025.
Copyright © 2025 Massachusetts Medical Society. All rights reserved, including those for text and data mining, AI training, and similar technologies.

timely. Whereas previous generations of AI tools to support clinical decision making have failed to augment human reasoning,[24] various studies have documented the expertlike performance of LLMs in several clinical reasoning competencies, including knowledge recall,[25,26] solution of complex diagnostic challenges,[27-29] probabilistic reasoning,[30] management reasoning,[31] and communication.[32] However, biased artifacts that reflect existing biases in health care[33] are likely to be incorporated during the training of AI models, a situation that has the potential to perpetuate (and inform[34]) diagnostic inequities.[35] Furthermore, LLMs confabulate[36] and exhibit cognitive biases that are similar to those of humans.[37] Therefore, although AI may serve as an adjunct, the final diagnosis and treatment plan must remain a human endeavor.[38,39] We focus this review on AI use in clinical reasoning, an area that carries high risk for learners and thus for their future patients[40] and must therefore be a priority for educators and learners.[41]

DESKILLING AND NEVER-SKILLING

As shown in Figure 2, using AI as a substitute for clinical reasoning (off-loading) rather than in support of clinical reasoning (informing) poses several risks with respect to skill development: deskilling,[42] never-skilling,[43] and mis-skilling. Overreliance on AI can lead to the loss of the clinical reasoning skills that the learner has just begun to acquire, including information recall, as shown with the use of Web searching.[44] In a study measuring the use of AI tools, cognitive off-loading, and critical thinking skills, researchers found a significant negative correlation between frequent use of AI tools and critical thinking abilities, which was mediated by increased

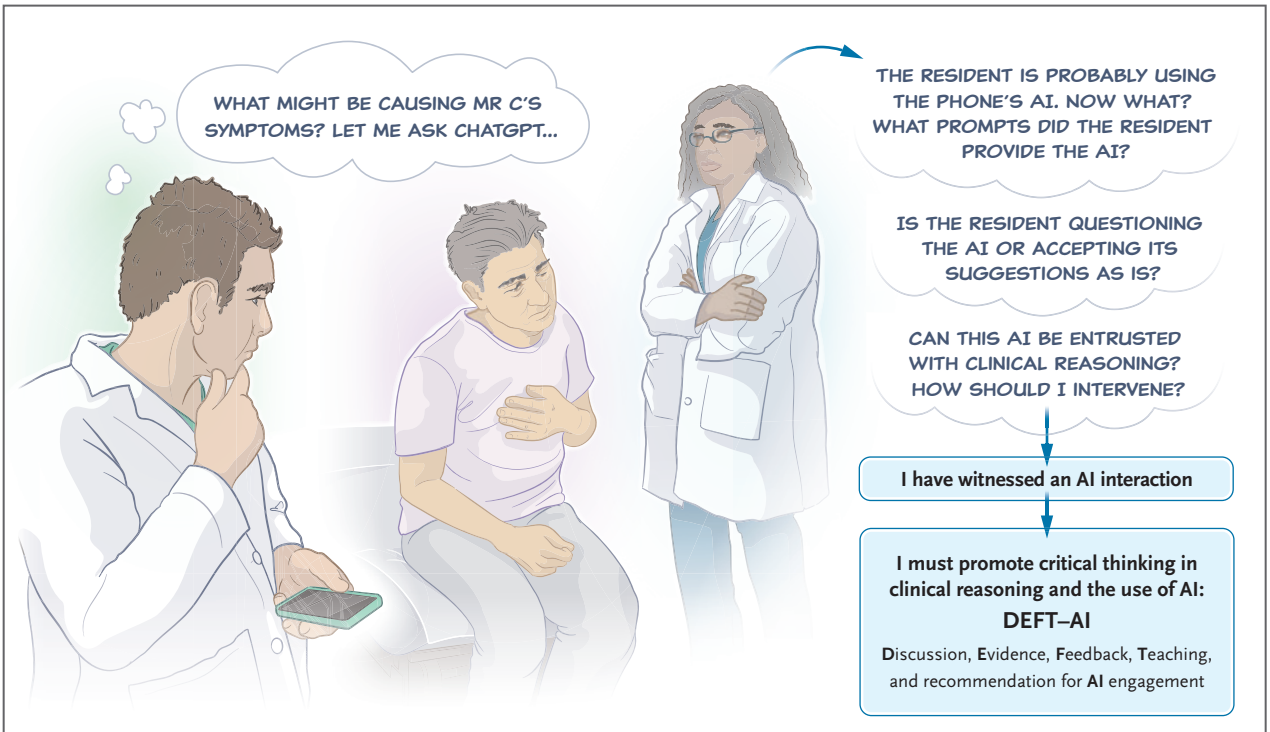

WHAT MIGHT BE CAUSING MR C'S SYMPTOMS? LET ME ASK CHATGPT...
THE RESIDENT IS PROBABLY USING THE PHONE'S AI. NOW WHAT? WHAT PROMPTS DID THE RESIDENT PROVIDE THE AI?
IS THE RESIDENT QUESTIONING THE AI OR ACCEPTING ITS SUGGESTIONS AS IS?
CAN THIS AI BE ENTRUSTED WITH CLINICAL REASONING? HOW SHOULD I INTERVENE?
I have witnessed an AI interaction
I must promote critical thinking in clinical reasoning and the use of AI:
DEFT–AI
Discussion, Evidence, Feedback, Teaching, and recommendation for AI engagement

Figure 1. An Educator Witnessing a Learner's Use of Artificial Intelligence (AI).

An educator, acting as a clinical preceptor, observes a resident who is using a large language model chatbot to assist with a differential diagnosis. The educator recognizes the inherent challenge of trusting an AI tool that may not be fully reliable. This moment of AI interaction prompts the educator to intervene in what could be a high-risk scenario for both the learner and the patient. By stepping in, the educator creates an opportunity to make critical thinking a scaffold and to foster deeper engagement with clinical reasoning and the responsible use of AI — an approach encapsulated in DEFT-AI (diagnosis, evidence, feedback, teaching, and recommendation for AI use).

The New England Journal of Medicine is produced by NEJM Group, a division of the Massachusetts Medical Society.
Downloaded from nejm.org by a a on August 20, 2025.
Copyright © 2025 Massachusetts Medical Society. All rights reserved, including those for text and data mining, AI training, and similar technologies.

cognitive off-loading.[45] Participants who reported higher reliance on AI showed reduced engagement in independent problem solving and analytic reasoning, findings that indicate a shift toward AI-generated solutions rather than personal cognitive effort. The same study showed that younger participants had greater dependence on AI tools and lower critical thinking scores than older participants,[45] which suggests that overuse of AI can result in a failure to develop critical thinking. These concerns are echoed in a randomized trial in which users who blindly adopted AI outputs without critical scrutiny performed more poorly with the use of AI than without it on tasks that required complex analytic skills,[5] an effect that was more pronounced among users with lower overall performance. The authors attributed this finding to "unengaged interaction with AI," with users bypassing their own judgment and acquiescing to the system's output.

MIS-SKILLING

Mis-skilling can occur when learners blindly entrust inaccurate or biased AI models with certain reasoning tasks.[9,33] For example, clinicians who were shown AI-generated diagnostic predictions with systematic biases, such as overestimating the likelihood of pneumonia in older patients or heart failure in patients with a high body-mass index, were more likely to adopt these incorrect predictions.[46] Furthermore, AI assistance disproportionately harmed clinicians with low

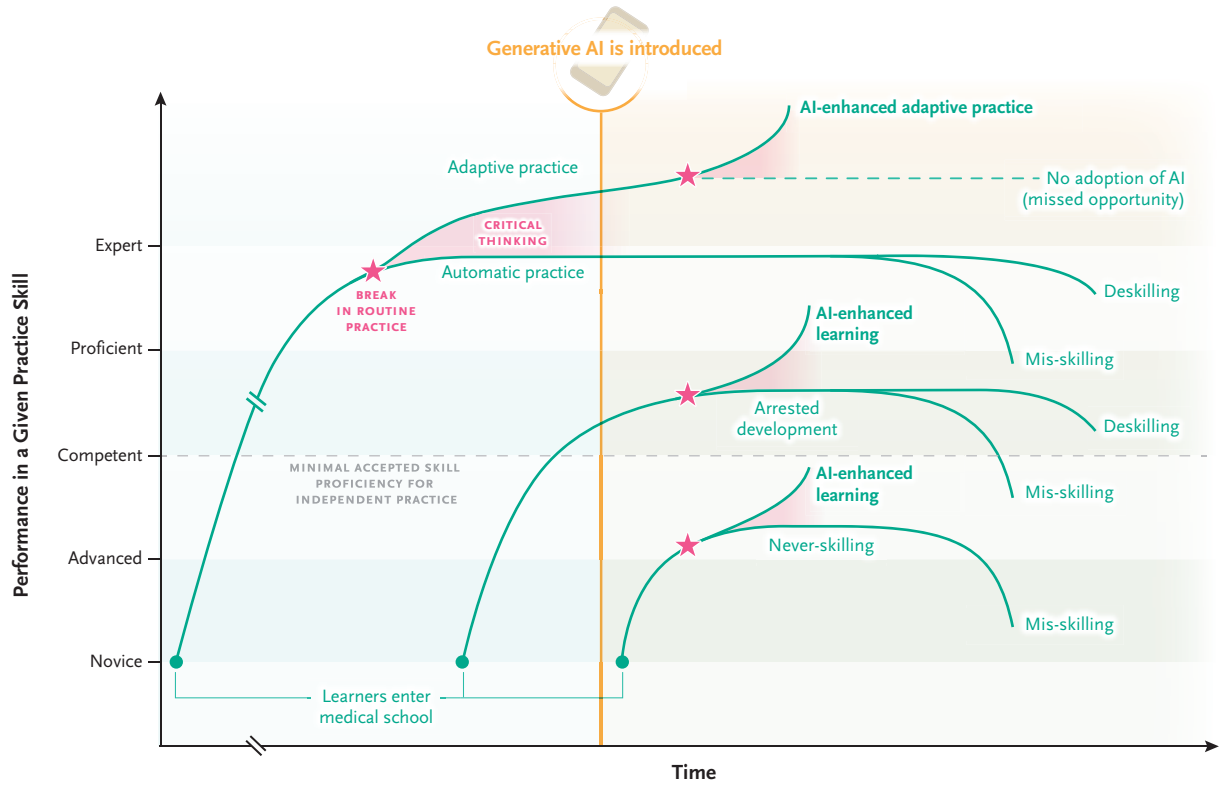


Figure 2. Development of Adaptive Practice and the Effects of AI.

Through practice and critical thinking, learners develop the ability to shift to innovative, adaptive practice in response to a break in routine, automatic practice (star). As they progress and enter clinical practice, the use of AI introduces both risks and opportunities. Cognitive off-loading onto AI can lead to overdependence on AI and "deskilling," whereas blind reliance on AI may result in "mis-skilling," with AI errors going unchallenged. If introduced too early, AI may prevent learners from acquiring essential skills ("never-skilling"). Conversely, judicious use of AI can enhance practice and learning by emphasizing the need for critical thinking and fostering effective human–AI collaboration.

The New England Journal of Medicine is produced by NEJM Group, a division of the Massachusetts Medical Society.
Downloaded from nejm.org by a a on August 20, 2025.
Copyright © 2025 Massachusetts Medical Society. All rights reserved, including those for text and data mining, AI training, and similar technologies.

baseline performance.[47,48] Indeed, when AI outperformed clinicians, the combination performed worse than AI alone. AI model explanations failed to mitigate these errors,[46,47] which may indicate that clinicians were unable to recognize and correct AI biases and that mis-skilling could be further reinforced. Ignoring AI recommendations, even when they are correct, highlights an underreliance on AI and missed opportunities for effective AI assistance. In contrast, when clinicians outperformed AI, the combination of human and AI reasoning increased performance, which suggests that high baseline performance promotes effective and safe AI assistance.[47]

EDUCATIONAL STRATEGIES DURING AI INTERACTIONS

Recognizing these risks, educational programs and institutions have established principles for the use of AI in medical education[15] and have begun to define competencies and curricula for the use of AI in health care.[19,49,50] However, educators still face the challenge of promoting the development of adaptive practice during in-the-moment AI interactions. Here, we propose a stepwise approach to learner–AI interactions that educators can use to model and scaffold critical thinking for the concurrent development of effective clinical skills and engagement with AI. As noted above, promoting strong foundational knowledge and skills maximizes the benefits and mitigates the risk of AI use and, we believe, must be a stated goal in any framework for effective and safe AI interactions. The surge of AI interactions must be framed as educational opportunities to increase AI and clinical literacy.[51]

DEFINING AI INTERACTIONS

The term "artificial intelligence" encompasses diverse definitions across disciplines and contexts, which can be grouped into technical, capability-based, and relational definitions (see Table S1 in the Supplementary Appendix, available with the full text of this article at NEJM.org). Of these three perspectives on AI, a relational definition of AI as illustrated in the resident–preceptor vignette (Fig. 1) is particularly helpful in a medical education context. This definition rests on the effects of AI on reasoning and practice rather than on its technical composition or capability.[52] An AI interaction is defined as a moment when, "in the context of a particular interaction, a computational artefact provides a judgement to inform an optimal course of action and . . . this judgement cannot be traced," which prompts the user to consider taking a leap of faith to trust the AI output.[53] Such leaps of faith are common when children first use a calculator or, in the case of AI in the clinic, when a medical student prompts generative AI for a differential diagnosis. The relational definition of AI is independent of the technological details of how the AI performs its task. Therefore, this definition of AI applies to the vignette whether the AI is a room-sized computer from the 1940s or a time-traveling robot from the future. Our point here is that the leap of faith inherent in AI interactions recognizes that AI-generated output cannot be fully trusted without verification,[54] which prompts the need to pause and critically assess the trustworthiness of the AI tool and its outputs.

CREATING AN EDUCATIONAL MOMENT

Once the educator recognizes an AI interaction, the opportunity arises to create an educational moment and cultivate critical thinking.[14] Building on existing models[55] that leverage the effects of a Socratic approach in enhancing critical thinking,[56] the DEFT (diagnosis, evidence, feedback, and teaching) framework emphasizes structured discussions of clinical reasoning, evidentiary support, and targeted feedback.[57] We propose an adapted approach, "DEFT-AI," which is designed to promote critical thinking and the development of adaptive practice when a learner, aided by AI, is engaged in clinical reasoning (Fig. 3 and Table S2). Although DEFT was developed outside the context of AI interactions, it has strong roots in educational theory,[55,57] and its common-sense approach will resonate with frontline educators negotiating any learner–AI interaction. Below, we review how each of the DEFT components leads to a recommendation for engaging with AI.

Diagnosis, Discussion, and Discourse

The educator begins by probing the learner's clinical reasoning process and use of AI. This step involves asking about the learner's synthesis of the clinical problem, which reflects data acquisition and inductive reasoning, and about the differential diagnosis, which reflects the learner's deductive reasoning and fund of knowledge. The educator also asks how the learner interacted

The New England Journal of Medicine is produced by NEJM Group, a division of the Massachusetts Medical Society.
Downloaded from nejm.org by a a on August 20, 2025.
Copyright © 2025 Massachusetts Medical Society. All rights reserved, including those for text and data mining, AI training, and similar technologies.

with AI: specifically, which AI tool and what prompts were used, whether and how follow-up prompts were structured to probe the validity of the AI-generated output, and whether the output influenced, replaced, or augmented the learner's diagnostic approach.

Evidence

At this stage, the educator probes the learner for the use of supporting and opposing evidence to evaluate the learner's medical and AI knowledge and the application of that knowledge. This process involves assessing the learner's use of diagnostic frameworks to generate diagnostic hypotheses, confirmation or refutation of these hypotheses, and the ability to leverage knowledge in order to generate alternative hypotheses — a hallmark of adaptive expertise. The educator may also ask the learner to justify the reasoning with supporting evidence and probe the learner's understanding of the pathobiology at play, application of relevant clinical guidelines and literature, and use of an evidence-based medicine framework.

Concurrently, the educator can engage the learner in self-assessment with respect to AI literacy. First, questions focus on technical understanding ("How do you think the AI reached its conclusions?"), critical appraisal ("What are the weaknesses of this AI? Can you evaluate the capabilities and limitations of the AI application?"), and practical application ("What problem is the AI you used designed to solve? Does this AI support or impede practice? What are effective prompting strategies?").[58,59] Next, educators should prompt learners to identify evidence supporting their use of the AI tool: "Which peer-reviewed research supports the use of this tool for clinical reasoning support?" General AI literacy scales can be used to ascertain the extent to which the learner understands the workings of the chosen AI application.[60] The educator can ask the learner to reason out loud through a modified case presentation without AI in order to assess the learner's ability to explore new problems independently of AI and to identify possible overreliance on AI.

Feedback

Guided self-reflection is central to this phase. The educator asks the learner to build on the self-evaluation and reflect on potential growth opportunities relevant to the case at hand and the learner's use of AI. These opportunities may include missed diagnostic considerations, gaps in medical knowledge relevant to the case,[61] and literacy in AI technology and its applications.

Teaching

The educator can build on the learner's self-reflection to provide feedback on the learner's reasoning performance and use of the specific AI tool in a particular activity, as well as general teaching principles tailored to the learner's needs. This may involve reinforcing the clinical reasoning process, encouraging the application of evidence-based medicine principles (including critical appraisal of the evidence), and promoting AI literacy (discussed below).

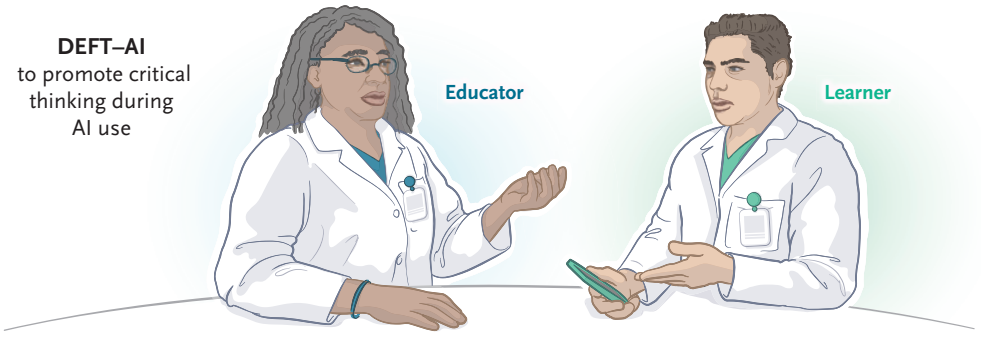
AI Engagement Recommendation

The educator concludes with recommendations that concurrently promote foundational skills and AI literacy. With rare exceptions, the educator should encourage ongoing practice with AI. More specifically, the educator may conclude that the learner can cautiously engage with the AI tool under indirect or intermittent direct supervision or that the learner can safely use the tool without supervision but with ongoing self-monitoring and education.

As this episode unfolds, one of two distinct human–AI collaboration behaviors is likely to emerge; these behaviors are known as centaur and cyborg[5] (Fig. 4). Centaur users strategically divide tasks between themselves and the AI; they allocate responsibilities on the basis of the strengths and capabilities of each entity, as would the mythical half-human, half-horse creature for which the behavior is named. AI may be used for ideation, summarization, or drafting, but centaur users rely on their clinical judgment for diagnosis and decision making. In contrast, cyborg users intertwine their work with AI across all stages of a task; the name derives from the cyborg in science fiction, which is a hybrid of human and machine. Cyborg users write an assessment plan with AI by iteratively prompting, correcting, and asking for justifications, then refining the output jointly with AI. This approach can be efficient and powerful, particularly for well-defined or low-risk tasks within the range of AI abilities, but carries the risk of never-skilling, deskilling, or mis-skilling due to automation bias.

The New England Journal of Medicine is produced by NEJM Group, a division of the Massachusetts Medical Society.
Downloaded from nejm.org by a a on August 20, 2025.
Copyright © 2025 Massachusetts Medical Society. All rights reserved, including those for text and data mining, AI training, and similar technologies.

DEFT–AI
to promote critical thinking during AI use



Diagnosis, **D**iscussion, and **D**iscourse	The educator asks for a description of the learner's specific use of AI.
What specific AI did you use?	**I used the free version of ChatGPT on my phone.**
How did you use AI in this process?	**I just typed in, "*What is the differential diagnosis for wheezing?*"**
What prompts did you enter in the app?	**I asked it for the best diagnostic test and treatment strategy.**
Evidence	The educator asks for an evaluation of the learner's evidence-based use of AI
How did you verify the AI-generated outputs?	**Hmm. I didn't. The answers seemed reasonable to me.**
Is the AI that you used shown to be accurate and safe?	**Yes. I keep seeing social media posts about how great it is at making diagnoses.**
Feedback	The educator asks the learner to reflect on growth opportunities in the use of AI.
How do you evaluate your own use of AI in this case?	**I think I've become quite familiar at using ChatGPT. I use it all the time now.**
How can you improve your use of AI?	**I can't wait for an AI that can interpret ECGs and chest radiographs. I should verify the AI outputs next time.**
Teaching	The educator provides focused teaching points based on findings from the conversation and recommends whether, when, and how to use AI safely moving forward.

Use AI tools that are known to be effective. Look for peer-reviewed evidence of their accuracy and safety. Our institution may have adapted and validated a similar model on the basis of high-quality data.

Prompting a chatbot is critical to generate valuable and accurate outputs. **Think of it as talking with a consultant:** provide enough specific information about the **Who** (the intended role of the AI and your role), the **Where** (description of the context), and the **What** (your goal and specific task or question). Always ask the AI to **explain its reasoning,** which improves its answers and lets you assess how it is thinking and how much to trust it. **One prompt is not enough:** have a conversation and give it feedback. Just like I did with you, you can also **ask it to engage in self-reflection and look for errors.**

AI is always prone to error and bias: always **verify and trust**. Make sure to check its answers against your knowledge, trusted sources of medical information, like publications from the NEJM Group, and your trusted peers, like me.

Recommendation for AI engagement — The educator provides learner-specific recommendations for the safe use of AI.

Keep practicing using AI to inform your reasoning rather than replace it. AI outputs are your preliminary inputs, just like a preliminary radiology report or automated ECG interpretation: verify, ***then*** trust. **Know when you can rely on it (cyborg) and when you need to confirm the outputs (centaur).**

The New England Journal of Medicine is produced by NEJM Group, a division of the Massachusetts Medical Society.
Downloaded from nejm.org by a a on August 20, 2025.
Copyright © 2025 Massachusetts Medical Society. All rights reserved, including those for text and data mining, AI training, and similar technologies.

Figure 3 (facing page). Use of DEFT-AI to Promote Critical Thinking during an AI Interaction.

After recognizing an AI interaction, the educator engages the learner in a structured educational moment to discuss the interaction, evaluate it, provide feedback, and teach clinical reasoning, as well as the use of AI. The discussion encompasses the learner's clinical reasoning process and approach to using AI, including the prompts used. The educator probes the learner for the use of supporting and opposing evidence to evaluate the learner's clinical and AI knowledge. This process helps determine whether the learner used AI to replace clinical reasoning or to inform it and offers a window into the learner's AI literacy. The educator then guides the learner in reflecting on growth opportunities and encourages critical thinking about AI interactions and clinical reasoning. Finally, the educator provides focused teaching on using AI effectively, selecting the right tools, and refining prompts. The learner is encouraged to adapt the use of AI to the task, switching between reliance on AI output (cyborg strategy) and confirmation of AI output (centaur strategy). Table S2 provides an expanded version of DEFT-AI.

Cyborg and centaur behaviors are not mutually exclusive. Users must shift between them on the basis of the task and the risks, an instance of adaptive practice that may maximize efficiency and innovation while preserving the development of clinical reasoning skills (Fig. 4).

Educators can help learners reflect on their interaction style with AI by using the cyborg–centaur framework. For example, educators can recommend that learners adopt a centaur approach to high-risk, uncertain, or diagnostic tasks, especially those outside the validated use of the AI tool (e.g., using a general-purpose AI chatbot such as ChatGPT for complex decision making). In these settings, educators could encourage learners to use AI and carefully validate its output. In contrast, cyborg strategies may be appropriate for low-risk, well-defined, or creative tasks for which the AI tool has demonstrated reliable performance, such as general communication drafts or initial ideation. In cyborg mode, learners can iteratively coconstruct solutions with AI, provided they maintain reflective oversight and can justify their approach to educators as needed.

Universally, educators should explicitly caution learners against passively adopting AI output without interrogation. Instead, adaptive engagement should be recommended: choosing the right interaction style for the right task, thinking critically about AI suggestions, refining prompts to obtain validated output, and modifying clinical interactions with AI as the learner's capabilities evolve and the context changes. By naming and discussing these interaction styles, educators help learners cultivate an adaptive practice with AI that can grow with technological advances and clinical complexity.

PROMOTING AI LITERACY

AI literacy begins with the ability to "call out" an AI interaction as a moment when one cannot retrace the judgment of the computer system and a cognitive pause is needed.[62] DEFT-AI promotes adaptive practice through critical thinking and facilitates seamless shifting among centaur, cyborg, and AI-independent behaviors, depending on the task. Two key areas of AI literacy for educators to focus on are a structured approach to critical appraisal of AI and effective use of prompts to maximize accuracy and relevance ("prompt engineering").

A critical appraisal of AI requires a structured evaluation of the tool itself and its output to make an evidence-based judgment of their trustworthiness. Evidence-based practice, rooted in the foundational principles of evidence-based medicine, offers a structured process to introduce new evidence into clinical practice.[63] Sackett's five-step model — asking the question, acquiring the best evidence, appraising the evidence, applying the findings to practice, and assessing outcomes — provides a comprehensive and structured framework for evidence-based practice.[64] Incorporating such a process can help determine the trustworthiness of AI in a particular task by structuring separate inquiries about the AI tool and its output.

Evidence-Based Evaluation of the AI Tool

To evaluate the trustworthiness of the AI tool itself, the first step is to determine the question that initiates the search for evidence. In the case of the vignette, the educator can ask the following question: "What is the accuracy of this LLM for differential diagnosis in adult ambulatory care patients?"

The second and third steps involve acquiring and appraising evidence, respectively, for the trustworthiness of AI, such as peer-reviewed research, AI scorecards[65] and related leaderboards (scoreboards that display the names and rankings of AI tools),[66,67] and regulatory information from

The New England Journal of Medicine is produced by NEJM Group, a division of the Massachusetts Medical Society.
Downloaded from nejm.org by a a on August 20, 2025.
Copyright © 2025 Massachusetts Medical Society. All rights reserved, including those for text and data mining, AI training, and similar technologies.

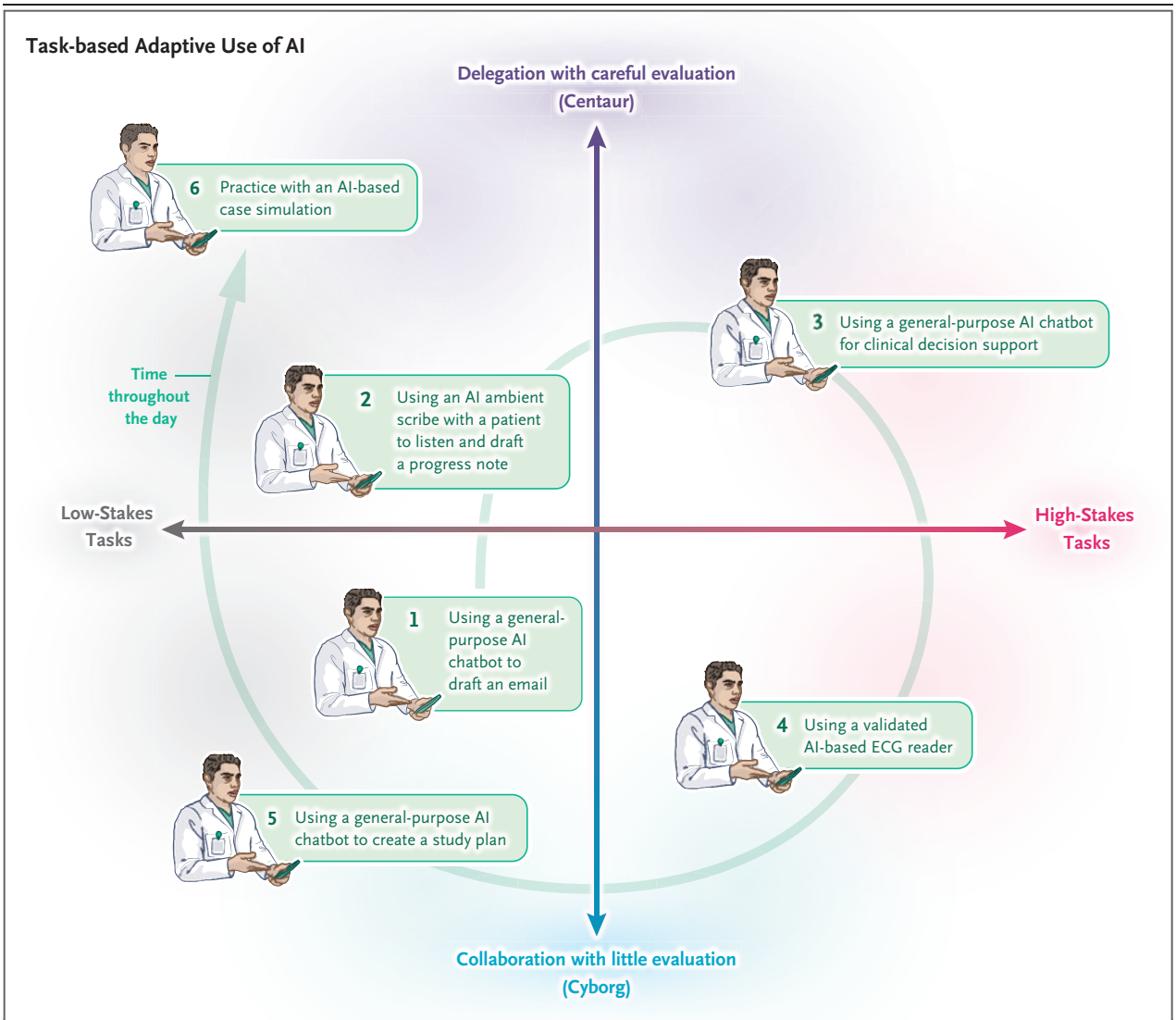

Task-based Adaptive Use of AI
Delegation with careful evaluation
(Centaur)
6 Practice with an AI-based case simulation
3 Using a general-purpose AI chatbot for clinical decision support
Time throughout the day
2 Using an AI ambient scribe with a patient to listen and draft a progress note
Low-Stakes Tasks
High-Stakes Tasks
1 Using a general-purpose AI chatbot to draft an email
4 Using a validated AI-based ECG reader
5 Using a general-purpose AI chatbot to create a study plan
Collaboration with little evaluation
(Cyborg)

Figure 4. Task-Based Adaptive Use of AI.

AI engagement behaviors should vary moment to moment, adapting to the type of task and its associated risk. The degree of human–AI integration can range from strategic delegation (centaur) to tight collaboration (cyborg). Centaur strategies are appropriate when human judgment must lead, particularly in high-stakes tasks or in situations in which the AI tool has not been validated for the specific intended use. With a centaur mindset, the clinician carefully delegates to AI and always evaluates the output. Cyborg strategies may enhance efficiency for low-risk, creative, or well-validated tasks. Adaptive AI use involves shifting between these modes on the basis of skill, task, and context.

health systems or government bodies (e.g., the Food and Drug Administration). Although scorecards, leaderboards, and regulatory frameworks are helpful in evaluating AI tools, they are insufficient for real-time judgment in educational moments and currently have limited usefulness for educators. It seems reasonable to assume that a comprehensive assessment of AI tools themselves is beyond the scope of most educators and learners.

Evidence-Based Evaluation of AI Output

Instead of assessing AI tools, clinicians are poised to assess AI output by integrating their clinical skills, patient preferences, and the research evidence to appraise the accuracy of the

The New England Journal of Medicine is produced by NEJM Group, a division of the Massachusetts Medical Society.
Downloaded from nejm.org by a a on August 20, 2025.
Copyright © 2025 Massachusetts Medical Society. All rights reserved, including those for text and data mining, AI training, and similar technologies.

output in a given clinical scenario.[64] Clinicians independently acquire and appraise evidence, such as established guidelines, published literature, or an expert consultant's opinion, about the clinical question that AI was used to answer and compare their conclusions with the AI output. A reliable concordance between AI and human outputs can decrease — but never remove — the need for human vigilance and promote trustworthiness in the use of the AI tool. Learners need to develop these independent skills so that they can reliably compare AI output with the output from their own clinical reasoning with regard to a clinical assessment of a patient.

Prompting LLMs

Effective prompting is a critical skill for maximizing the usefulness of LLM-enabled medical applications and performing meaningful evaluation of them.[68] Like a request for a consultant's opinion about a patient, the LLM prompt determines the relevance and quality of the response. As in human consultations, vague or poorly framed queries can lead to confusion or misdirection. The output is only as good as the input.

Key features of a good prompt include specificity and context provision. A well-defined query, as compared with a general query, yields a more precise response (e.g., "What are the most common risk factors for coronary artery disease?" instead of "Tell me about heart disease"). Context provision helps LLMs generate more relevant outputs by incorporating background information (e.g., "I'm a pulmonologist caring for a patient in my clinic with asthma that is refractory to inhaler therapy. List several hypotheses to explain the lack of treatment response."). Biased prompts that lead the LLM (e.g., "What is the diagnosis? I think it is asthma.") can promote "sycophantic" responses[69] and cognitive biases similar to those of humans.[37] Therefore, practice and iteration with unbiased prompts during conversations with LLMs are essential.[54]

Providing example cases as part of the initial prompt improves accuracy. In addition, asking the AI model to "think out loud" (i.e., chain-of-thought prompting) reveals the reasoning discourse of the AI, which enhances the accuracy of its outputs[70,71] and allows assessment of the reasoning.[72] For example, the prompt to "generate a prioritized differential diagnosis" is followed by "explain your reasoning." Newer LLM models have chain-of-thought reasoning embedded in their interface,[71] a feature that facilitates critical appraisal of their outputs. When an AI output seems inaccurate or sparks further reflection, engaging the model in a follow-up conversation — such as prompting it to explain or revise its response — can transform passive use into active learning, strengthen critical thinking, and maximize the educational value of AI.

VERIFY AND TRUST

Despite the technical advancements of AI tools, their use still requires leaps of faith with careful consideration; the need for verification is at the heart of AI interactions. As medical learners increasingly use these tools, often as an integral part of their patient evaluations, medical educators must face the reality that AI interactions are here to stay. Although critical thinking is the bulwark against the deskilling, never-skilling, and mis-skilling that can arise from an overreliance on AI,[73] the opportunity to promote critical thinking as a scaffold can accelerate the development of adaptive practice skills and concurrently improve the AI literacy of both learners and educators. DEFT-AI provides a structured and common-sense approach to promote critical thinking during learner–AI interactions and underscores the importance of establishing the validity of an AI output as part of the AI use process. The onus lies with educators to inculcate in their trainees the conviction that verification is key to AI use. To do this effectively will require curricular redesign, with close collaboration among AI developers, health care systems, and educational programs, in order to promote AI competencies among learners and educators. We must also include systematic assessment of learner–AI interactions in the educational settings in which they occur.[74] Without governance structures, rigorous validation frameworks, and ongoing monitoring, the risk of AI-driven errors and biases may outweigh the benefits of AI technologies, which may thus jeopardize medical education rather than improve it. Ultimately, fostering a "verify and trust" paradigm is crucial for ensuring that AI is a beneficial augmentation of human expertise.

The New England Journal of Medicine is produced by NEJM Group, a division of the Massachusetts Medical Society.
Downloaded from nejm.org by a a on August 20, 2025.
Copyright © 2025 Massachusetts Medical Society. All rights reserved, including those for text and data mining, AI training, and similar technologies.

Disclosure forms provided by the authors are available with the full text of this article at NEJM.org.

We thank Drs. Sanjay Desai and Richard Schwartzstein for their thoughtful feedback on an earlier version of the manuscript. Figure 2 is inspired by the work of Ericsson[75] and Pusic et al.[12]

Jeffrey M. Drazen, M.D., Raja-Elie Abdulnour, M.D., and Judith L. Bowen, M.D., Ph.D., are the editors of the Medical Education series.

AUTHOR INFORMATION

[1]Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston; [2]Department of Medical Education and Pediatrics, University of Illinois College of Medicine, Chicago; [3]Department of Medicine, University of California, San Francisco, San Francisco; [4]Department of Anesthesia and Perioperative Care, University of California, San Francisco, San Francisco.

REFERENCES

1. Pusic MV, Birnbaum RJ, Thoma B, et al. Frameworks for integrating learning analytics with the electronic health record. J Contin Educ Health Prof 2023;43:52-9.

2. Younas A, Zeng Y. A philosophical inquiry into AI-inclusive epistemology. August 30, 2024 (https://dx.doi.org/10.2139/ssrn.4902415). preprint.

3. Cooper A, Rodman A. AI and medical education — a 21st-century Pandora's box. N Engl J Med 2023;389:385-7.

4. Topol EJ. Deep medicine: how artificial intelligence can make healthcare human again. New York: Basic Books, 2019.

5. Dell'Acqua F, McFowland E III, Mollick E, et al. Navigating the jagged technological frontier: field experimental evidence of the effects of AI on knowledge worker productivity and quality. Boston: Harvard Business School, September 22, 2023 (https://www.hbs.edu/ris/Publication%20Files/24-013_d9b45b68-9e74-42d6-a1c6-c72fb70c7282.pdf).

6. Lavigne E, Lopez A, Frandon J, et al. AI-standardized clinical examination training on OSCE performance. NEJM AI 2025;2(8) (https://ai.nejm.org/doi/10.1056/AIoa2500066).

7. Boscardin CK, Gin B, Golde PB, Hauer KE. Chatgpt and generative artificial intelligence for medical education: potential impact and opportunity. Acad Med 2024;99:22-7.

8. Risko EF, Gilbert SJ. Cognitive offloading. Trends Cogn Sci 2016;20:676-88.

9. Schwartzstein RM. Clinical reasoning and artificial intelligence: can AI really think? Trans Am Clin Climatol Assoc 2024;134:133-45.

10. Coiera E, Fraile-Navarro D. AI as an ecosystem — ensuring generative AI is safe and effective. NEJM AI 2024;1(9) (https://ai.nejm.org/doi/10.1056/AIp2400611).

11. Waldock WJ, Lam G, Baptista A, Walls R, Sam AH. Which curriculum components do medical students find most helpful for evaluating AI outputs? BMC Med Educ 2025;25:195.

12. Pusic MV, Santen SA, Dekhtyar M, et al. Learning to balance efficiency and innovation for optimal adaptive expertise. Med Teach 2018;40:820-7.

13. Schumacher DJ, Englander R, Carraccio C. Developing the master learner: applying learning theory to the learner, the teacher, and the learning environment. Acad Med 2013;88:1635-45.

14. Richards JB, Hayes MM, Schwartzstein RM. Teaching clinical reasoning and critical thinking: from cognitive theory to practical application. Chest 2020;158:1617-28.

15. Association of American Medical Colleges. Principles for the responsible use of artificial intelligence in and for medical education (https://www.aamc.org/about-us/mission-areas/medical-education/principles-ai-use).

16. Furfaro D, Celi LA, Schwartzstein RM. Artificial intelligence in medical education: a long way to go. Chest 2024;165:771-4.

17. Clay MA II, Sikon AL, Lypson ML, et al. Teaching while learning while practicing: reframing faculty development for the patient-centered medical home. Acad Med 2013;88:1215-9.

18. Ilgen JS, Teunissen PW, de Bruin ABH, Bowen JL, Regehr G. Warning bells: How clinicians leverage their discomfort to manage moments of uncertainty. Med Educ 2021;55:233-41.

19. Car J, Ong QC, Erlikh Fox T, et al. The Digital Health Competencies in Medical Education framework: an international consensus statement based on a Delphi Study. JAMA Netw Open 2025;8(1):e2453131.

20. Al-Akchar M, Ibrahim AM, Salih M, et al. Learning electrocardiogram interpretation — insights from residents and a proposed solution in an observational study. J Comm Med Pub Health Rep 2021;2:1-7 (https://acquaintpublications.com/get/1-JCMPHR2021093006%20Revised_Galley_Proof-1635917652.pdf).

21. Rajpurkar P, Lungren MP. The current and future state of AI interpretation of medical images. N Engl J Med 2023;388:1981-90.

22. Tu T, Azizi S, Driess D, et al. Towards generalist biomedical AI. NEJM AI 2024;1(3) (https://ai.nejm.org/doi/10.1056/AIoa2300138).

23. Tierney AA, Gayre G, Hoberman B, et al. Ambient artificial intelligence scribes to alleviate the burden of clinical documentation. NEJM Catal 2024;5(3) (https://catalyst.nejm.org/doi/full/10.1056/CAT.23.0404).

24. Howell MD, Corrado GS, DeSalvo KB. Three epochs of artificial intelligence in health care. JAMA 2024;331:242-4.

25. Katz U, Cohen E, Shachar E, et al. GPT versus resident physicians — a benchmark based on official board scores. NEJM AI 2024;1(5) (https://ai.nejm.org/doi/10.1056/AIdbp2300192).

26. Singhal K, Tu T, Gottweis J, et al. Toward expert-level medical question answering with large language models. Nat Med 2025;31:943-50.

27. Kanjee Z, Crowe B, Rodman A. Accuracy of a generative artificial intelligence model in a complex diagnostic challenge. JAMA 2023;330:78-80.

28. Goh E, Gallo R, Hom J, et al. Large language model influence on diagnostic reasoning: a randomized clinical trial. JAMA Netw Open 2024;7(10):e2440969.

29. McDuff D, Schaekermann M, Tu T, et al. Towards accurate differential diagnosis with large language models. Nature 2025;642:451-7.

30. Rodman A, Buckley TA, Manrai AK, Morgan DJ. Artificial intelligence vs clinician performance in estimating probabilities of diagnoses before and after testing. JAMA Netw Open 2023;6(12):e2347075.

31. Goh E, Gallo RJ, Strong E, et al. GPT-4 assistance for improvement of physician performance on patient care tasks: a randomized controlled trial. Nat Med 2025;31:1233-8.

32. Tu T, Schaekermann M, Palepu A, et al. Towards conversational diagnostic artificial intelligence. Nature 2025;642:442-50.

33. Zack T, Lehman E, Suzgun M, et al. Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: a model evaluation study. Lancet Digit Health 2024;6(1):e12-e22.

34. Ferryman K, Mackintosh M, Ghassemi M. Considering biased data as informative artifacts in AI-assisted health care. N Engl J Med 2023;389:833-8.

35. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science 2019;366:447-53.

36. Hatem R, Simmons B, Thornton JE. Chatbot confabulations are not hallucinations. JAMA Intern Med 2023;183:1177.

The New England Journal of Medicine is produced by NEJM Group, a division of the Massachusetts Medical Society.
Downloaded from nejm.org by a a on August 20, 2025.
Copyright © 2025 Massachusetts Medical Society. All rights reserved, including those for text and data mining, AI training, and similar technologies.

37. Wang J, Redelmeier DA. Cognitive biases and artificial intelligence. NEJM AI 2024;1(12) (https://ai.nejm.org/doi/full/10.1056/AIcs2400639).
38. Verghese A, Shah NH, Harrington RA. What this computer needs is a physician: humanism and artificial intelligence. JAMA 2018;319:19-20.
39. Restrepo D, Rodman A, Abdulnour R-E. Conversations on reasoning: large language models in diagnosis. J Hosp Med 2024;19:731-5.
40. Newman-Toker DE, Nassery N, Schaffer AC, et al. Burden of serious harms from diagnostic error in the USA. BMJ Qual Saf 2024;33:109-20.
41. Rencic J, Trowbridge RL Jr, Fagan M, Szauter K, Durning S. Clinical reasoning education at US medical schools: results from a national survey of internal medicine clerkship directors. J Gen Intern Med 2017;32:1242-6.
42. Aquino YSJ, Rogers WA, Braunack-Mayer A, et al. Utopia versus dystopia: professional perspectives on the impact of healthcare artificial intelligence on clinical roles and skills. Int J Med Inform 2023;169:104903.
43. Rafel JB. Proceedings and abstracts of the 2024 Artificial Intelligence and Medical Education Macy Conference, November 18, 2024. Atlanta: Josiah Macy Jr. Foundation, 2024.
44. Sparrow B, Liu J, Wegner DM. Google effects on memory: cognitive consequences of having information at our fingertips. Science 2011;333:776-8.
45. Gerlich M. AI tools in society: impacts on cognitive offloading and the future of critical thinking. Societies (Basel) 2025;15:6 (https://www.mdpi.com/2075-4698/15/1/6).
46. Jabbour S, Fouhey D, Shepard S, et al. Measuring the impact of AI in the diagnosis of hospitalized patients: a randomized clinical vignette survey study. JAMA 2023;330:2275-84.
47. Yu F, Moehring A, Banerjee O, Salz T, Agarwal N, Rajpurkar P. Heterogeneity and predictors of the effects of AI assistance on radiologists. Nat Med 2024;30:837-49.
48. Vaccaro M, Almaatouq A, Malone T. When combinations of humans and AI are useful: a systematic review and meta-analysis. Nat Hum Behav 2024;8:2293-303.
49. Russell RG, Lovett Novak L, Patel M, et al. Competencies for the use of artificial intelligence-based tools by health care professionals. Acad Med 2023;98:348-56.
50. Tolentino R, Baradaran A, Gore G, Pluye P, Abbasgholizadeh-Rahimi S. Curriculum frameworks and educational programs in AI for medical students, residents, and practicing physicians: scoping review. JMIR Med Educ 2024;10:e54793.
51. Chen JH. Who's training whom? NEJM AI 2024;1(5) (https://ai.nejm.org/doi/10.1056/AIp2400006).
52. Bearman M, Ajjawi R. Learning to work with the black box: pedagogy for a world with artificial intelligence. Br J Educ Technol 2023;54:1160-73 (https://bera-journals.onlinelibrary.wiley.com/doi/10.1111/bjet.13337).
53. Bearman M, Ajjawi R. When I say… artificial intelligence. Med Educ 2024;58:1273-5.
54. Zwaan L. Cognitive bias in large language models: implications for research and practice. NEJM AI 2024;1(12) (https://ai.nejm.org/doi/full/10.1056/AIe2400961).
55. Neher JO, Gordon KC, Meyer B, Stevens N. A five-step "microskills" model of clinical teaching. J Am Board Fam Pract 1992;5:419-24.
56. Jantausch BA, Bost JE, Bhansali P, Hefter Y, Greenberg I, Goldman E. Assessing trainee critical thinking skills using a novel interactive online learning tool. Med Educ Online 2023;28:2178871.
57. Savaria MC, Min S, Aghagoli G, Tunkel AR, Hirsh DA, Michelow IC. Enhancing the one-minute preceptor method for clinical teaching with a DEFT approach. Int J Infect Dis 2022;115:149-53.
58. Laupichler MC, Aster A, Haverkamp N, Raupach T. Development of the "scale for the assessment of non-experts' AI literacy" — an exploratory factor analysis. Comput Hum Behav Rep 2023;12:100338 (https://www.sciencedirect.com/science/article/pii/S2451958823000714?via%3Dihub).
59. Wang B, Rau P-LP, Yuan T. Measuring user competence in using artificial intelligence: validity and reliability of artificial intelligence literacy scale. Behav Inf Technol 2022;42:1324-37 (https://www.tandfonline.com/doi/full/10.1080/0144929X.2022.2072768).
60. Lintner T. A systematic review of AI literacy scales. NPJ Sci Learn 2024;9:50.
61. Singh H. Editorial: helping health care organizations to define diagnostic errors as missed opportunities in diagnosis. Jt Comm J Qual Patient Saf 2014;40:99-101.
62. Lee JY, Szulewski A, Young JQ, Donkers J, Jarodzka H, van Merriënboer JJG. The medical pause: Importance, processes and training. Med Educ 2021;55:1152-60.
63. Dusin J, Melanson A, Mische-Lawson L. Evidence-based practice models and frameworks in the healthcare setting: a scoping review. BMJ Open 2023;13(5):e071188.
64. Sackett DL. Evidence-based medicine. Semin Perinatol 1997;21:3-5.
65. Mitchell M, Wu S, Zaldivar A, et al. Model cards for model reporting. In: Proceedings and abstracts of the Conference on Fairness, Accountability, and Transparency, January 29–31, 2019. New York: Association for Computing Machinery, 2019.
66. Bommasani R, Liang P, Lee T. Holistic evaluation of language models. Ann N Y Acad Sci 2023;1525:140-6.
67. Stanford Center for Research on Foundation Models. HELM lite — leaderboard: core scenarios. 2024 (https://crfm.stanford.edu/helm/lite/latest/#/leaderboard).
68. Meskó B. Prompt engineering as an important emerging skill for medical professionals: tutorial. J Med Internet Res 2023;25:e50638.
69. Sharma M, Tong M, Korbak T, et al. Towards understanding sycophancy in language models. October 20, 2023 (https://doi.org/10.48550/arXiv.2310.13548) preprint.
70. Wei J, Wang X, Schuurmans D, et al. Chain-of-thought prompting elicits reasoning in large language models. January 28, 2022 (https://doi.org/10.48550/arXiv.2201.11903). preprint.
71. Brodeur PG, Buckley TA, Kanjee Z, et al. Superhuman performance of a large language model on the reasoning tasks of a physician. December 14, 2024 (https://doi.org/10.48550/arXiv.2412.10849). preprint.
72. Cabral S, Restrepo D, Kanjee Z, et al. Clinical reasoning of a generative artificial intelligence model compared with physicians. JAMA Intern Med 2024;184:581-3.
73. Adler-Milstein J, Redelmeier DA, Wachter RM. The limits of clinician vigilance as an AI safety bulwark. JAMA 2024;331:1173-4.
74. Vokinger KN, Soled DR, Abdulnour R-EE. Regulation of AI: learnings from medical education. NEJM AI 2025;2(5) (https://ai.nejm.org/doi/10.1056/AIp2401059).
75. Ericsson KA. Deliberate practice and the acquisition and maintenance of expert performance in medicine and related domains. Acad Med 2004;79:Suppl:S70-S81.
Copyright © 2025 Massachusetts Medical Society.

The New England Journal of Medicine is produced by NEJM Group, a division of the Massachusetts Medical Society.
Downloaded from nejm.org by a a on August 20, 2025.
Copyright © 2025 Massachusetts Medical Society. All rights reserved, including those for text and data mining, AI training, and similar technologies.